MICROBIAL GENOMICS

RESEARCH ARTICLE
Sidorczuk *et al.*, *Microbial Genomics* 2023;9:000999
DOI 10.1099/mgen.0.000999

MICROBIOLOGY SOCIETY
OPEN DATA OPEN ACCESS

Genomic characterization of enterohaemolysin-encoding haemolytic *Escherichia coli* of animal and human origin

Katarzyna Sidorczuk[1], Adrianna Aleksandrowicz[2], Michał Burdukiewicz[3,4], Robert A. Kingsley[5] and Rafał Kolenda[2,5,]*

Abstract

Enterohaemolysin (Ehx) and alpha-haemolysin are virulence-associated factors (VAFs) causing the haemolytic phenotype in *Escherichia coli*. It has been shown that chromosomally and plasmid-encoded alpha-haemolysin are characteristic of specific pathotypes, virulence-associated factors and hosts. However, the prevalence of alpha- and enterohaemolysin does not overlap in the majority of pathotypes. Therefore, this study focuses on the characterization of the haemolytic *E. coli* population associated with multiple pathotypes in human and animal infectious diseases. Using a genomics approach, we investigated characteristic features of the enterohaemolysin-encoding strains to identify factors differentiating enterohaemolysin-positive from alpha-haemolysin-positive *E. coli* populations. To shed light on the functionality of Ehx subtypes, we analysed Ehx-coding genes and inferred EhxA phylogeny. The two haemolysins are associated with a different repertoire of adhesins, iron acquisition or toxin systems. Alpha-haemolysin is predominantly found in uropathogenic *E. coli* (UPEC) and predicted to be chromosomally encoded, or nonpathogenic and undetermined *E. coli* pathotypes and typically predicted to be plasmid-encoded. Enterohaemolysin is mainly associated with Shiga toxin-producing *E. coli* (STEC) and enterohaemorrhagic *E. coli* (EHEC) and predicted to be plasmid-encoded. Both types of haemolysin are found in atypical enteropathogenic *E. coli* (aEPEC). Moreover, we identified a new EhxA subtype present exclusively in genomes with VAFs characteristic of nonpathogenic *E. coli*. This study reveals a complex relationship between haemolytic *E. coli* of diverse pathotypes, providing a framework for understanding the potential role of haemolysin in pathogenesis.

DATA SUMMARY

The authors confirm all supporting data, code and protocols have been provided within the article or through supplementary data files.

(1) Supplementary data containing all supplementary figures and tables (Supplementary Material).
(2) Dataset S1: list of genomes used in this study (Supplementary Dataset S1).
(3) Supplementary dataset 2: results of mapping of the contigs identified as plasmids by Platon to 306 reference plasmids from PLSDB containing enterohaemolysin (Supplementary Dataset S2).

INTRODUCTION

Escherichia coli is a versatile bacterial species encompassing commensals and pathogens of animals and humans. The latter are highly adapted strains with acquired virulence attributes and infect hosts effectively, causing a wide range of intestinal and extraintestinal infections worldwide [1, 2]. Pathogenic *E. coli* are classified into intestinal (InPEC) and extraintestinal (ExPEC) based on the site of infection. Due to the characteristic virulence-associated factors/genes (VAFs/VAGs) and shared patterns of pathogenicity for particular hosts, they are further categorized into pathotypes (reviewed in detail in [2]). InPECs

Received 07 September 2022; Accepted 24 February 2023; Published 27 April 2023
Author affiliations: [1]Department of Bioinformatics and Genomics, Faculty of Biotechnology, University of Wrocław, Wrocław, Poland; [2]Department of Biochemistry and Molecular Biology, Faculty of Veterinary Medicine, Wrocław University of Environmental and Life Sciences, Wrocław, Poland; [3]Clinical Research Centre, Medical University of Białystok, Białystok, Poland; [4]Autonomous University of Barcelona, Institute of Biotechnology and Biomedicine, Cerdanyola del Vallès, Spain; [5]Quadram Institute Biosciences, Norwich Research Park, Norwich, UK.
*Correspondence: Rafał Kolenda, Rafal.Kolenda@quadram.ac.uk
Keywords: *Escherichia coli*; haemolytic; enterohaemolysin; alpha-haemolysin; adhesins; iron acquisition; toxin; EhxA.
Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files.Nine supplementary figures, four supplementary tables and two supplementary data files are available with the online version of this article.

Impact Statement

Enterohaemolysin and alpha-haemolysin belong to the family of RTX toxins that cause lysis of erythrocytes. Haemolysis on blood agar is used in bacteriology as part of routine diagnostics. Due to phenotypic and genotypic similarities, enterohaemolysin and alpha-haemolysin are often mistaken for each other. In this study, we characterize the haemolytic *E. coli* population of multiple pathotypes involved in human and animal infectious diseases. We show that these haemolysins are associated with a different repertoire of adhesins, iron acquisition or toxin systems. Alpha-haemolysin occurs in UPEC, nonpathogenic or undetermined *E. coli* pathotypes and may be chromosomally or plasmid-encoded. Enterohaemolysin is typically associated with STEC and EHEC and predicted to be plasmid-encoded. The only pathotype where both types of haemolysin are found is aEPEC. Moreover, we identify a new subtype of EhxA in genomes with a nonpathogenic-like virulence-associated factor profile and suggest an association between EhxA sequence variation and altered *E. coli* toxicity towards the host. These insights reveal a complex relationship between haemolytic *E. coli* of diverse pathotypes, providing a framework for understanding the potential role of haemolysin in pathogenesis.

include enteropathogenic (EPEC), enterotoxigenic (ETEC), enteroaggregative (EAEC), enteroinvasive (EIEC), diffusely adherent (DAEC) and Shiga toxin-producing (STEC) *E. coli*, whereas ExPEC are categorized into uropathogenic (UPEC), sepsis-associated (SEPEC), neonatal meningitis (NMEC) and avian pathogenic (APEC) *E. coli* strains [1, 3].

The virulence of microorganisms, including *E. coli*, is a multistep interaction between bacterial and host cells involving numerous VAFs, comprising both extra- and intracellular structures [4]. The ability of bacteria to colonize host tissue is determined by VAFs such as adhesins, iron acquisition systems and toxins [5]. The primary role of adhesins is binding to epithelial cells, resulting in invasion and ultimately leading to host colonization. Some may also contribute to the interactions between bacteria and abiotic surfaces during biofilm formation enabling survival in diverse conditions and environments [6, 7]. The current classification system distinguishes fimbrial and nonfimbrial adhesins [8]. However, the adhesion process may also involve structures and mechanisms indirectly contributing to the binding and colonization. They are known as atypical adhesins and include flagella, lipopolysaccharide, and type III and VI secretion systems (T3SS, T6SS) [9].

The expression of genes encoding iron acquisition and storage systems allows bacteria to survive in iron-deficient environments [10]. The most relevant and effective mechanisms to acquire iron during infection include the production of siderophores, iron transporters and haem/hemin uptake systems [11]. Alternatively, bacteria secrete haemolysins to lyse red blood cells and release haemoglobin, and/or produce haemoglobin proteases that make the iron available [12].

Three types of haemolysin have been described in *E. coli*, alpha-haemolysin (α-Hly), enterohaemolysin (Ehx/EHEC-hly) and silent haemolysin (HlyE) [13]. α-Hly is encoded by an operon consisting of four genes – *hlyCABD*. The *hlyA* gene encodes a functional cytotoxic haemolysin following activation by the product of *hlyC*. In turn, products of *hlyB* and *hlyD*, along with the outer-membrane channel TolC, are involved in the secretion of haemolysin through the bacterial cell envelope [14]. α-Hly exhibits cytotoxic activity against various types of cells, including erythrocytes, monocytes, granulocytes, endothelial and epithelial cells. Ehx is encoded on a plasmid by an operon consisting of *ehxC*, *ehxA*, *ehxB* and *exhD* genes with functions analogous to *hlyC*, *hlyA*, *hlyB* and *hlyD*, respectively [15]. Ehx targets erythrocytes, which after lysis release haemoglobin and haem, positively influencing *E. coli* proliferation and growth. The role of haemolysis in *E. coli* pathogenicity and its association with other virulence factors and disease outcome remains inconclusive. We recently reported the association of VAF repertoire with the context of the *hlyCABD* cluster in the genome and demonstrated that chromosomally and plasmid-encoded α-Hly are characteristic of specific pathotypes and hosts [13]. Chromosomally encoded α-Hly was associated with UPEC strains from human infections, whereas plasmid-encoded α-Hly was found in nonpathogenic *E. coli* and less often in various intestinal pathotypes isolated from animals. Moreover, haemolytic *E. coli* with plasmid-encoded α-Hly showed a similar adhesin profile to nonpathogenic strains, whereas chromosome-encoded α-Hly had a high prevalence of Auf, P and S fimbriae characteristic for UPEC [13].

Based on our previously published analysis, we tested the hypothesis that genome context of Ehx genes is related to the virulence factor repertoire of Ehx-positive *E. coli*, as was observed for α-Hly-positive *E. coli*. Moreover, we tested whether Ehx-positive strains possess genomic signatures not found in α-Hly-positive *E. coli* to provide crucial information for the differentiation of function of Ehx and α-Hly. Using a genomics approach, we investigated characteristic features of the genome of Ehx-encoding *E. coli* isolated from animals and humans to provide a one-health perspective on these pathogens. Finally, to shed light on the prevalence of Ehx variants and subtypes, we performed an in-depth analysis of Ehx coding genes and EhxA phylogeny.

METHODS

Genome acquisition and analysis

A RefSeq collection of *E. coli* genomes was screened for the presence of enterohaemolysin-encoding genes *ehxCABD* (from reference genome EDL 933; GenBank accession number X86087.1), yielding 2399 enterohaemolysin-positive genomes (Supplementary Dataset S1, available in the online version of this article). Additionally, the collection of 1122 *hlyCABD*-positive (alpha-haemolysin-positive) *E. coli* genomes was obtained from Kolenda *et al*. [13] to allow comparative studies (Supplementary Dataset S2). The genomic context of alpha-haemolysin-coding genes was determined as in Kolenda *et al*. [13]. Clermont typing was performed with the use of clermonTyping to determine the phylogroup of *E. coli* under analysis [16]. Multilocus sequence typing (MLST) was determined with the PubMLST database and mlst software [17]. *In silico* serotyping was performed with ABRicate using the EcOH database [18]. Bayesian analysis of population structure (BAPS) was carried out by fastBAPS with use of optimized symmetric prior to hyperparameter optimization [19]. Pathotypes were predicted for *ehxCABD*-positive *E. coli* genomes by identification of VAG profile with BLAST, as listed in Table S1 [13]. Genomes that did not fit the criteria for any pathotype were classified as unknown (NA). The genomic context of enterohaemolysin-coding genes was analysed using BLAST by comparing reference sequences of 5′-upstream of *ehxC* and 3′-downstream of *ehxD* (Table S2). Reference sequences of 5′-upstream of *ehxC* and 3′-downstream of *ehxD* specific only for plasmid-encoded enterohaemolysin were selected by analysis with use of complete chromosome and plasmid sequences in the Nucleotide collection (nr/nt) database using BLAST. Platon software with default settings was used as additional confirmation that *ehxCABD* operon is encoded on a plasmid [20]. First, contigs that encode plasmid sequences were identified with Platon and next they were analysed with BLAST for the presence of *ehxCABD* (from reference genome EDL 933; GenBank accession number X86087.1). Additionally, contigs identified as plasmids/part of a plasmid with Platon were searched against 306 plasmids encoding *ehxCABD* identified in PLSDB (described in paragraph 'Analysis of plasmids'). All BLAST results with identity and query coverage equal to or above 95% were considered positive. The pangenome of *ehxCABD*-positive *E. coli* was analysed via Roary. Variable sites were extracted from core gene alignment generated in Roary with SNP-sites and used in phylogenetic analysis with FastTree 2.1 [21, 22]. The phylogenetic tree was annotated by iTOL [23].

Comparison of gene frequency between *hlyCABD*- and *ehxCABD*-positive *E. coli*

The pangenome of *hlyCABD*- and *ehxCABD*-positive *E. coli* was determined with the use of Roary [24]. Genes in the pangenome were functionally annotated by using eggnog-mapper [25]. Gene frequencies between *hlyCABD*- and *ehxCABD*-positive *E. coli* were compared with Scoary [26]. Clusters of orthologous groups (COGs) were compared between *hlyCABD*- and *ehxCABD*-positive *E. coli* for all genes with a Bonferroni *P*-value <0.001 in Scoary analysis. Comparison between numbers of genes belonging to the same functional group between *hlyCABD*- and *ehxCABD*-positive *E. coli* was performed with the chi-squared test of independence implemented in R stats package [27].

Adhesin, toxin and iron acquisition gene frequency determination

E. coli genomes were tested for adhesins, toxins and iron acquisition genes as published in Kolenda *et al*. [13] using 567 genes coding for adhesins, toxins and iron acquisition collected from GenBank database. Gene frequency in analysed genomes was tested with ABRicate. The genes from the same system or with similar functionality were presented as one VAF, as shown in Table S3. System prevalence was compared with chi-squared test of independence implemented in the R stats package.

Analysis of plasmids

A plasmid database, PLSDB, was queried for the presence of alpha-haemolysin or enterohaemolysin genes [28]. Sequences of *hlyCABD* were acquired from *E. coli* strain UTI89 (GenBank accession number CP000243.1) and sequences of *ehxCABD* genes were obtained from *E. coli* strain EDL933 (GenBank accession number X86087.1). Pangenomes for plasmid sequences were determined with Panaroo [29]. Functional annotation of plasmid genes was performed with the eggnog-mapper [25]. Comparison between numbers of genes belonging to the same functional group in haemolysin was performed with the Kruskal–Wallis test. Gene presence and absence was utilized in hierarchical clustering of plasmids with use of hclust from R package stats.

Analysis of EhxCABD cluster

Sequences of *ehxCABD* genes were extracted via BLAST and the sequence of *E. coli* strain EDL933 (GenBank accession number X86087.1) was used as reference. Next, the nucleotide sequences were translated by UGENE [30]. Variant clustering was carried out by CD-HIT [31]. Protein sequences were considered the same variant if they were 100% identical and had the same length. Protein functionality was assessed based on protein length and all variants with a sequence at least 1% shorter than the most frequent variant were considered non-functional. EhxA phylogeny was performed using RAxML and employing a PROTGAMMADUMMY2 substitution model and 500 bootstrap replications [32]. The EhxA tree was annotated with iTOL [23]. Subtyping of *ehxA* gene was based on TaqI restriction digestion of *ehxA* gene sequence as described by Lorenz *et al*. [33]. TaqI restriction digestion of *ehxA* gene sequences was performed *in silico* with the use of DigestDNA function from DECIPHER R package [34].

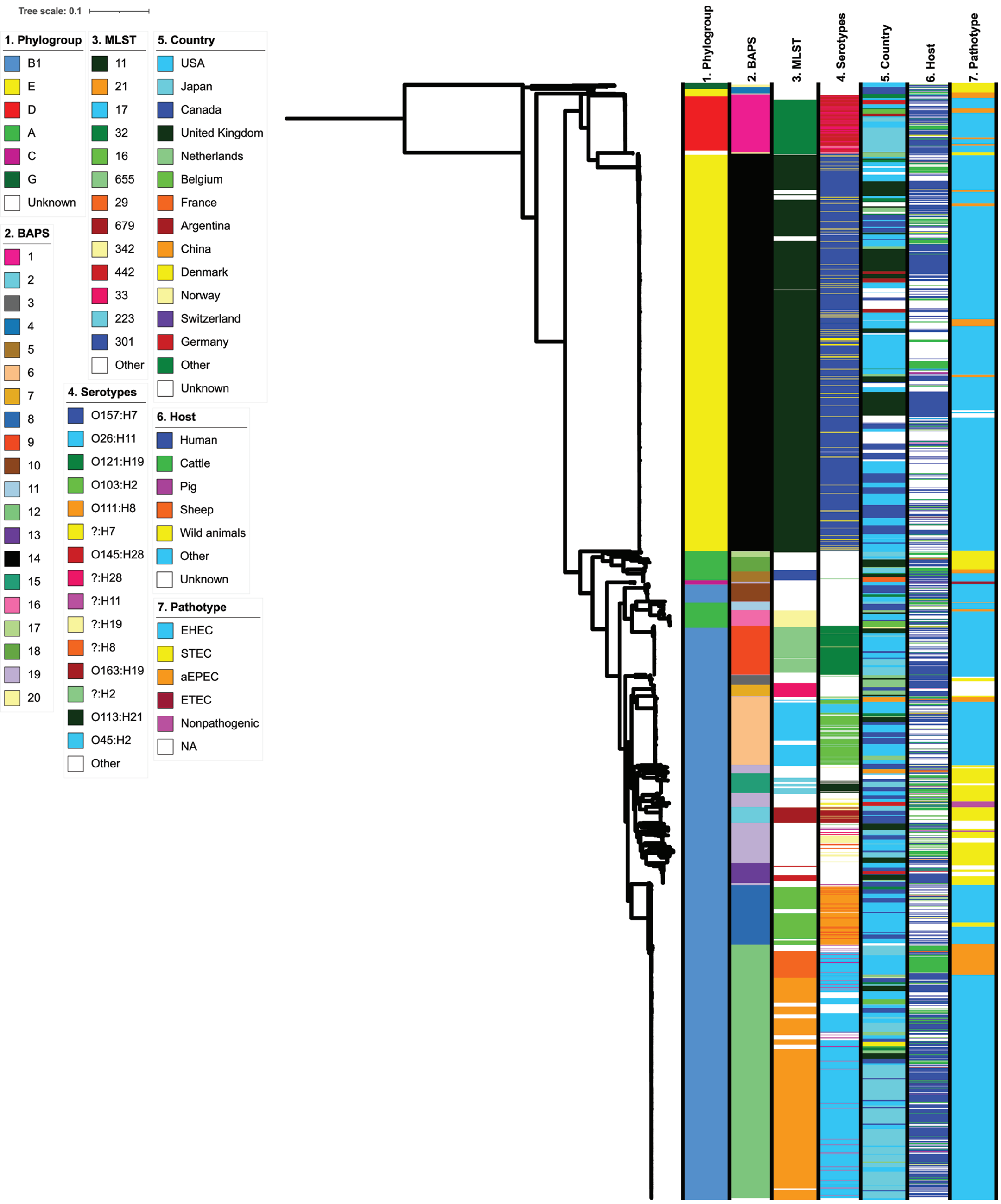

Fig. 1. Phylogenetic relationship of 2399 genomes of enterohaemolytic *E. coli*. Phylogroup, BAPS group, MLST, serotype, origin, host and pathotype have been annotated on the midpoint-rooted tree using iTOL.

RESULTS

Characterization of enterohaemolysin-positive *E. coli* genomes

RefSeq database contained 2399 *E. coli* genomes positive for 4 enterohaemolysin genes, which were categorized into 6 phylogroups (Fig. 1). More than half (52%) of the analysed strains belonged to the B1 phylogroup, whereas phylogroup E had 36% isolates. Phylogroups D, A, C and G were much less prevalent, consisting of 5.35, 4.96, 0.17 and 0.17% of strains, respectively (Fig. 1). Investigation of the genetic structure with BAPS revealed 20 clusters, 11 of which belonged to phylogroup B1, whereas E1 and D phylogroups were nearly homogenous, represented mainly by clusters 14 and 1, respectively (Figs 1 and S1a). *In silico* prediction of the serotype formula identified 136 serotypes, of which only 15 were represented by 20 or more genomes (Figs 1 and S1c). The strains were isolated from diverse geographical locations, including 31 countries across 6 continents, although the majority (84%) were isolated in the USA (34.4%), Japan (18.1%), Canada (17.9%) and the UK (13.4%) (Figs 1 and S1b). Considering

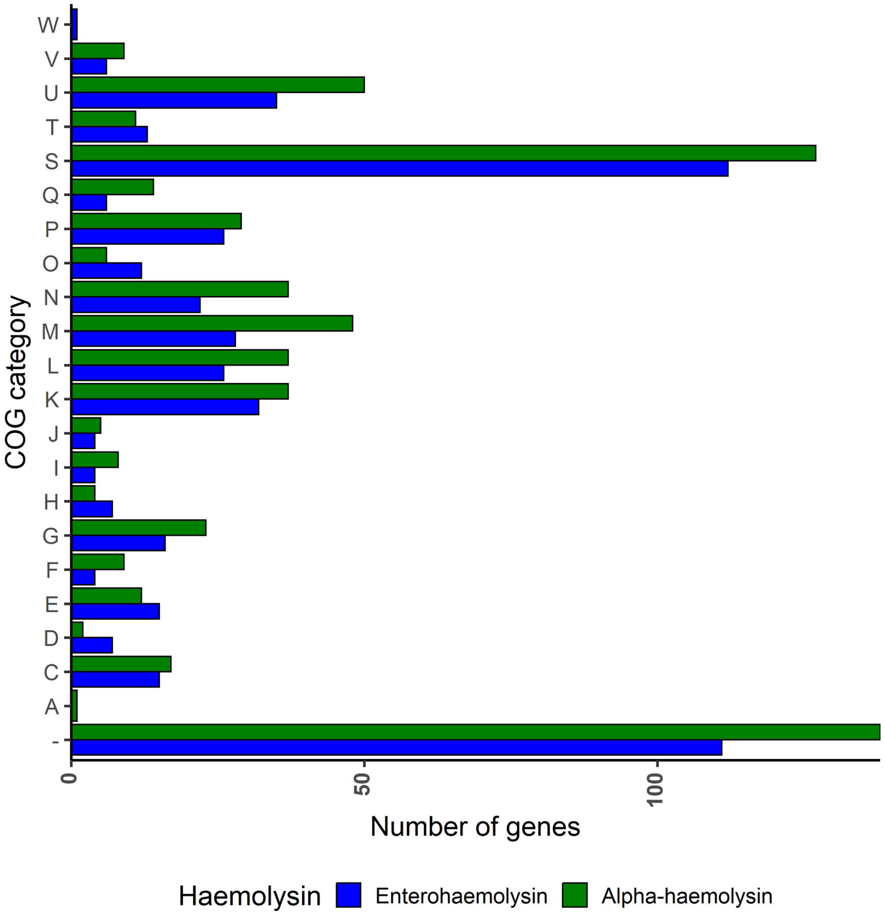

Fig. 2. Functional annotation of genes with different prevalence in alpha-haemolysin- and enterohaemolysin-positive *E. coli*. Barplot with gene frequency in clusters of orthologous groups of alpha-haemolysin- and enterohaemolysin-positive *E. coli*. Genes with a different frequency between *E. coli* with alpha-haemolysin and enterohaemolysin were functionally annotated with eggnog-mapper. The number of genes from a particular COG is shown on the *x*-axis. COG groups are shown on the *y*-axis. Bar colours represent different haemolysins and are described in the legend below the plot.

the isolation source, they were obtained mainly from humans and cattle (Figs 1 and S1e). The enterohaemolysin-positive *E. coli* strains were of 149 STs, highlighting the widespread distribution of enterohaemolysin genes within *E. coli* (Figs 1 and S1f). Pathotype prediction revealed that most strains were represented by EHEC (77.9%), followed by STEC (11.9%) and atypical enteropathogenic *E. coli* (aEPEC) (6.3%) (Fig. 1). The least prevalent groups were nonpathogenic *E. coli* and ETEC, with 0.7 and 0.04% of the total, respectively. Analysis of the genomic context of enterohaemolysin clusters by comparing reference sequences of 5′-upstream of *ehxC* and 3′-downstream of *ehxD* suggests that it can only be found on plasmids, but the context of *ehxCABD* could not be determined for 17.7% of genomes with that method (Fig. S2). Additional analysis using Platon software revealed that the majority of contigs (from 95% of strains) that contain *ehxCABD* are identified as plasmid sequences. Mapping of the contigs identified as plasmids by Platon to 306 reference plasmids from PLSDB showed that contigs from 84.7% of genomes can be aligned to at least 1 of the plasmids (Supplementary Dataset S2).

Comparison of enterohaemolysin- and alpha-haemolysin-positive *E. coli*

In order to obtain a comprehensive overview of alpha-haemolysin- and enterohaemolysin-positive *E. coli* populations, their genomes were compared with respect to phylogroups, BAPS clusters, STs and serotypes. All isolates were present in one of 6 phylogroups – A, B1, B2, C, D and E – and further subdivided into 29 clusters defined by hierarchical Bayesian analysis (Fig. S3a). The majority of *hlyCABD*-positive isolates (65%) belonged to B2, while more than half of *ehxCABD*-positive strains represented the B1 phylogroup (52%). Both types of haemolysins displayed a comparable number of STs and serotypes (Fig. S3c, e), indicating their similar genetic and antigenic diversity. Twenty-four of 275 STs were prevalent in both groups and comprised 1102 isolates in total. Twenty-two out of 303 serotypes were commonly identified in 901 genomes from both haemolysin groups. *HlyCABD*-positive genomes deposited in the RefSeq database were more diverse in terms of country of isolation than *ehxCABD*-positive genomes (Fig. S3b). The investigations of genome context of *hlyCABD* and *ehxCABD* suggested that most alpha-haemolytic

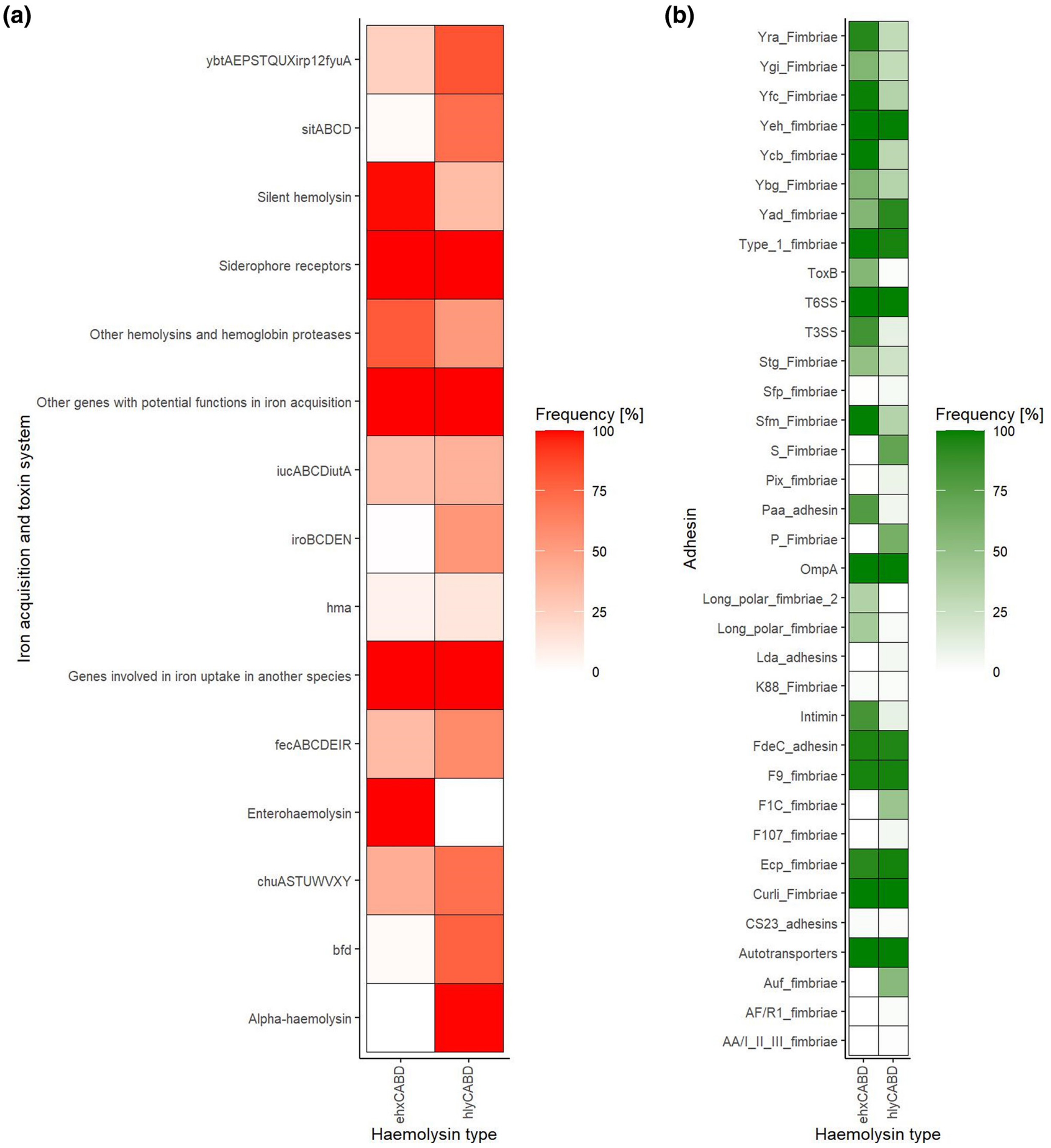

Fig. 3. Frequency of toxin, iron acquisition and adhesin systems in alpha-haemolytic and enterohaemolytic *E. coli*. The frequency of 128 toxins and iron acquisition and 284 adhesin genes providing information about the presence of 15 toxins or iron acquisition systems and 50 adhesins or adhesion-related molecules was tested in 1122 alpha-haemolytic and 2399 enterohaemolytic *E coli*. Genome context is shown on the *x*-axis. Iron acquisition and toxin systems (a) or adhesins (b) are listed on the *y*-axis. The colour gradient is proportional to the frequency of each system, and the colour scale is shown on the legends attached to heatmaps. In the case of adhesins (b), each system that had a prevalence lower than or equal to 1% in both groups was not included in the plot.

isolates from the B2 and D phylogroup had alpha-haemolysin encoded on chromosomes (Fig. S3d). Regarding the B1 and E phylogroup, more than 70% of enterohaemolysins and alpha-haemolysins were predicted to be plasmid-encoded, similarly to alpha-haemolysins from these groups. In phylogroup A, enterohaemolysins were predicted to be plasmid-encoded, whereas alpha-haemolysins as encoded on a chromosome or plasmid. In the case of group C, the determination of *ehxCABD* context was not achieved. Comparison of genome context with the host origin of strains revealed that animal isolates mainly possessed

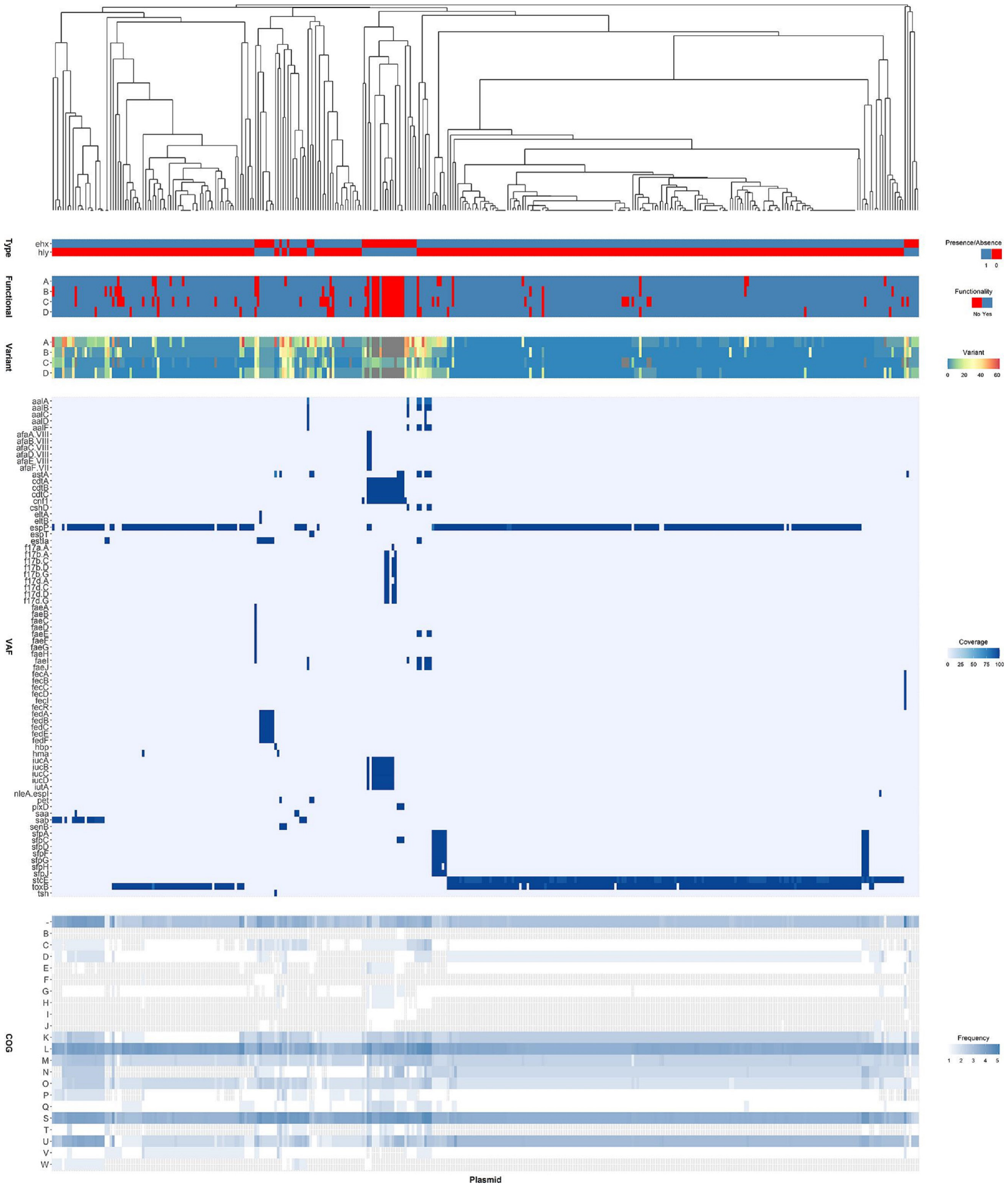

Fig. 4. Comparison of alpha-haemolysin- and enterohaemolysin-bearing plasmids. Analysis of gene presence, haemolysin variation and functionality, VAFs and COGs in alpha-haemolysin- and enterohaemolysin-bearing plasmids. Sequences of 41 alpha-haemolysin- and 306 enterohaemolysin-bearing plasmids were used to generate the pangenome with Panaroo. Next, a gene presence/absence tree was generated and annotated with: haemolysin type (ehx, enterohaemolysin; hly, alpha-haemolysin), haemolysin protein variation and functionality (A, EhxA or HlyA; B, EhxB or HlyB; C, EhxC or HlyC; D, EhxD or EhxD), VFDB VAF coverage and COG frequency. For each heatmap, the plasmid is set on the *x*-axis and haemolysin type, haemolysin protein, VAFs or COGs are shown on the *y*-axes. Legends with colour scales are provided on the right side of each heatmap.

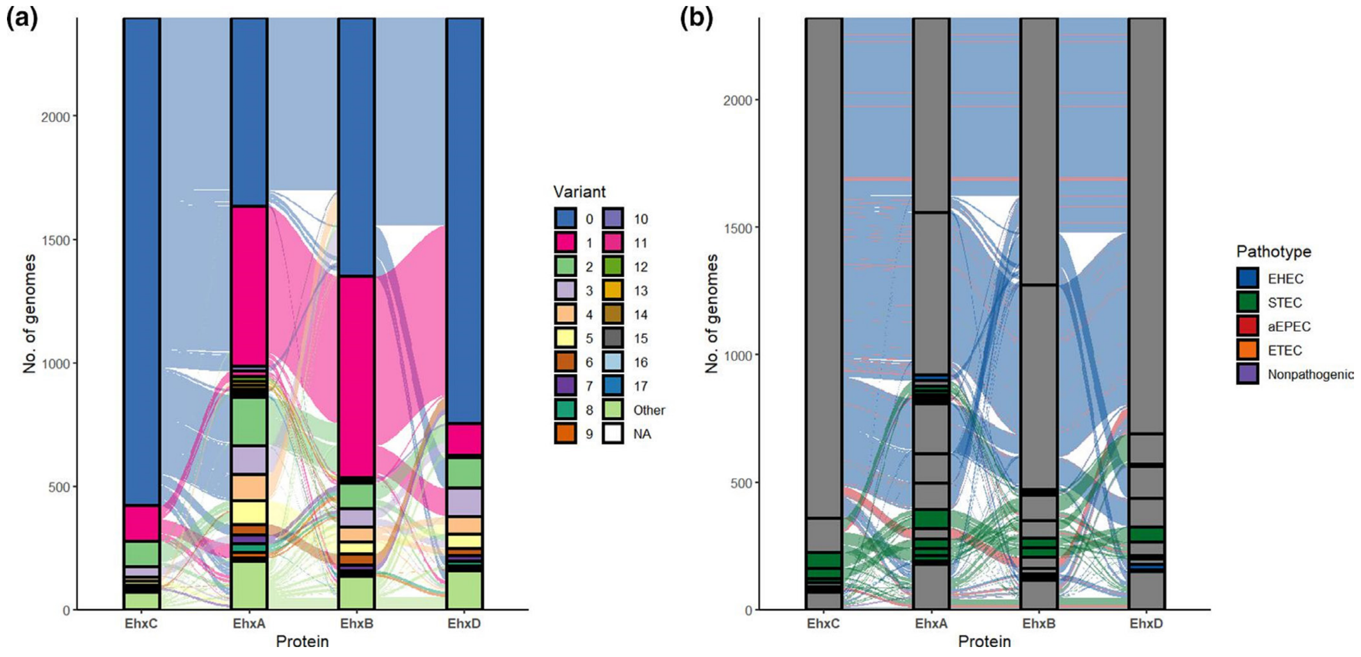

Fig. 5. EhxCABD variants in enterohaemolysin-positive *E. coli*. Analysis of EhxCABD variant prevalence in enterohaemolysin-positive *E. coli* integrated with information about *E. coli* pathotype. The number of genomes is shown on the *y*-axis and the names of the proteins are shown on the *x*-axis. Colour streams and bars represent protein variants (a) or pathotypes (b) and are shown separately on the right side of each alluvial plot. In plot b) the bar has a grey colour if one variant is present in more than one pathotype. Variants present in fewer than 10 isolates were regrouped as 'Other' in plot (a) (the plot with all variants is shown in Fig. S8a). All genomes with undetermined pathotypes are not shown on plot (b) (the plot with all genomes is shown in Fig. S8c).

plasmid-encoded haemolysins of both types, whereas chromosomally encoded toxins were considerably more common (76%) for alpha-haemolytic human isolates (Fig. S3g).

The pangenome of *hlyCABD*- and *ehxCABD*-positive *E. coli* was determined and used to annotate gene function classification associated with strains encoding each haemolysin. Based on inferred gene function, all genes with different frequencies in *hlyCABD*- and *ehxCABD*-positive *E. coli* were divided into 22 functional groups. Differences in gene frequency between alpha-haemolysin- and enterohaemolysin-positive *E. coli* were found in five groups, unassigned (–), M, N, Q and U (Fig. 2). Of note, these COGs were identified more frequently in *hlyCABD*-positive strains. The largest group (unassigned) covered genes whose function was not found in the eggNOG database. This group had more than 10% higher prevalence in alpha-haemolytic strains compared to enterohaemolytic strains (*P*<0.005, chi-squared test). The second group (Q) with the highest differences in frequency between both *E. coli* groups contained genes associated with secondary metabolite biosynthesis or transport and catabolism processes (*P*<0.05, chi-squared test, Fig. S4). Additional subgroups of gene function were distinguished in group Q, with most genes encoding proteins possessing oxidoreductase activity and/or belonging to the short-chain dehydrogenases/reductases (SDRs) family. Their functions also included amino acid polymerization and lipoteichoic acid biosynthesis. Moreover, alpha-haemolytic strains were characterized by the higher prevalence of genes responsible for intracellular trafficking, secretion and vesicular transport (U, *P*<0.05, chi-squared test), encoding chaperone–usher proteins, fimbrial and fimbrial-like adhesins, and proteins associated with type II and III secretion systems (Fig. S4). Genes with similar functions were found in COG N, where adhesin-encoding proteins were identified (*P*<0.01, chi-squared test, Fig. S4). Another functional class (M) included genes constituting a component of cell wall and outer membrane structures and biogenesis, i.e. various enzymes, transporters, chaperone–usher and outer-membrane proteins (*P*<0.005, chi-squared test, Fig. S4).

Alpha-haemolysin- and enterohaemolysin-positive *E. coli* differ in profiles of adhesin, toxin and iron acquisition systems frequency

Since our comparison of predicted gene functional classes associated with haemolysins revealed differences in adhesion and toxin secretion functions (Figs 2 and S4), we next investigated the prevalence of adhesins, toxins and iron acquisition systems using a manually curated set of genes with confirmed functions. Overall, 65 VAFs encoded by 412 VAGs were identified in at least 1 isolate. The VAFs consisted of 50 adhesins and 15 toxin or iron acquisition system genes (Fig. 3). However, the prevalence between alpha-haemolytic and enterohaemolytic *E. coli* differed for 33 VAFs (*P*<0.05, chi-squared test). *E. coli*

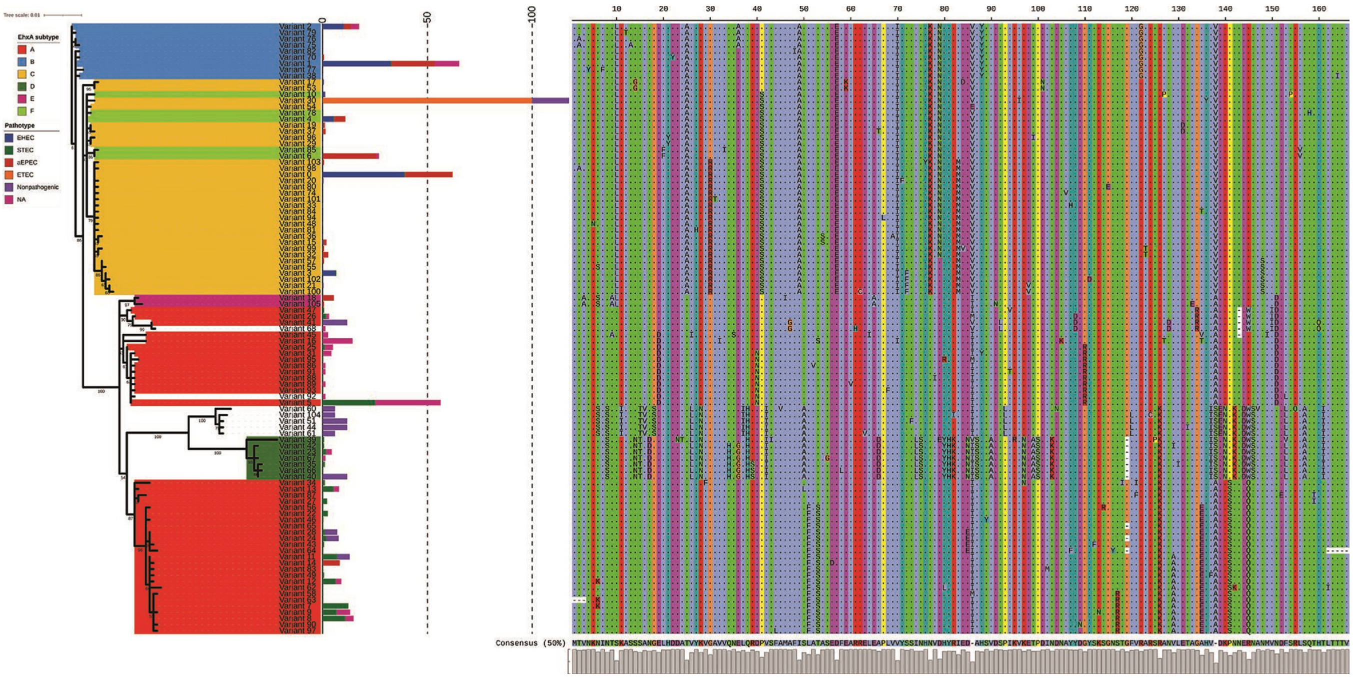

Fig. 6. EhxA sequence variation. Phylogenetic analysis of 99 variants of EhxA found in 2399 enterohaemolysin-positive *E. coli* genomes. Relative prevalence of each variant in various pathotypes (as bars), amino acid-variable sites (166) and EhxA subtypes (colours of the clades) were annotated with use of iTOL. Amino acid-variable sites numbers on the plot do not reflect actual site numbers in the whole EhxA protein alignment. Colour keys for pathotypes shown on the barplot and EhxA subtypes marked on the phylogenetic tree are shown on the legends 'Pathotype' and 'EhxA subtype', respectively.

from both groups had genes encoding a silent haemolysin with different prevalence, 34.6% among *hlyCABD*-positive strains and 99.3% in *ehxCABD*-positive strains. Two VAGs associated with iron acquisition and storage systems, i.e. *pic* and *espC* belonging to the group 'Other haemolysins and hemoglobin proteases', were present more frequently in *hlyCABD*-positive isolates ($P<0.001$, chi-squared test). Regarding adhesins, Auf, S, P, F1C, Yad, Pix, AF/RI, Sfp and Ecp fimbriae had a higher prevalence in *hlyCABD*-positive strains ($P<0.001$, chi-squared test). Enterohaemolytic strains had a higher prevalence of intimin, Paa, Sfm, T3SS, Ycb, Yfc, Yra, ToxB, Lpf, Lpf2, Ygi, Stg and Ybg fimbriae ($P<0.001$, chi-squared test). The frequency of 29 out of 65 VAFs was similar in all isolates. Three of them were associated with iron acquisition systems and included agents involved in iron uptake in other bacterial species, other factors with potential functions in iron acquisition and siderophore receptors. The remaining 26 VAFs encoded fimbrial and nonfimbrial adhesins, including autotransporters group. Interestingly, the investigation of single genes encoding autotransporters revealed differences in the prevalence of *ehaA*, *ehaG*, *ehaB*, *espP*, *cah*, *ypjA*, *yeeJ*, *ycgV* and *flu* (Fig. S5, $P<0.05$, chi-squared test). Only the *flu* gene that encoded an autotransporter was found more frequently in alpha-haemolytic *E. coli*.

High genetic diversity defines plasmids bearing alpha-haemolysin and enterohaemolysin

Considering that both haemolysins can be located on the plasmids, the diversity of these DNA molecules was assessed. All plasmids that contained genes from analysed haemolysins were downloaded from the PLSDB plasmid database. There were 41 and 306 plasmids that encoded at least 1 gene from the alpha- or enterohaemolysin operon, respectively. Most plasmids were circular (Fig. S6c). Their length and GC content varied from 44632 to 242187 bp and from 43.2 to 52.4%, respectively (Fig. S6d, e). Plasmid typing with PlasmidFinder and pMLST revealed that alpha- and enterohaemolysins are mainly found on different plasmid types (Fig. S6a, b), with both haemolysins only occurring in six PlasmidFinder groups (Fig. S6a). Analysis of pangenomes for alpha- and enterohaemolysin-bearing plasmids revealed high diversity in gene content among both groups. The pangenome of alpha-haemolysin-positive plasmids contained 1234 genes, similarly to enterohaemolysin-positive plasmids, which encoded 1160 genes. For most plasmids, the only genes common in the pangenomes were the haemolysin-coding genes. Functional annotation and analysis of plasmid genes revealed differences in the frequency of genes belonging to 15 functional groups between alpha-haemolysin and enterohaemolysin-positive plasmids (Figs 4 and S7, Table S4). Moreover, clustering based on gene presence/absence showed that plasmids from these haemolysin types grouped separately, which was associated with differences in VAF prevalence (Fig. 4). Additionally, gene presence/absence was also associated with haemolysin protein variations and functionality. Up to three alpha-haemolysin genes (*hlyABD*) were lost in 13 plasmids that clustered together, whereas the *ehxC* gene from enterohaemolysin was not found in 14 plasmids. Furthermore, the investigation of gene functionality, i.e. if it encodes a full-length protein without frameshifts, identified that 17 and 57 (41.5 and 18.6%, respectively) plasmids with alpha-haemolysin and enterohaemolysin encoded potentially non-functional haemolysins. Taken together, the data indicate that alpha-haemolysin and enterohaemolysin can be encoded on heterogeneous groups of plasmids, which share similar characteristics dependent on the haemolysin type they encode.

Global diversity of enterohaemolysin reveals pathotype-associated variants and subtypes

To assess the diversity and possible contribution of EhxCABD sequence variation to the enterohaemolysin function, the enterohaemolysin coding sequences were analysed. The highest number of variants was found for EhxA (106) followed by EhxD (84), EhxB (80) and EhxC (50) (Figs 5a and S8a). In some cases, potential loss of function mutations were detected. The highest numbers of these mutations were present in EhxD and EhxC, totalling 42 and 35, respectively, followed by only 9 strains with non-functional EhxA or EhxB. Overall, 97 of 2399 strains encoded potentially non-functional enterohaemolysin. When protein variants were contextualized with information about pathotypes from which the variants originated, it was apparent that most variants were present in more than one pathotype. It was interesting to note that EHEC and aEPEC possessed the same variants. This was also true for STEC, nonpathogenic *E. coli* and isolates with unidentified pathotypes as shown in Figs 5b, S8b and S8c. The variants found exclusively in one pathotype were characteristic of STEC.

Since EhxA functions as a toxin and also had the highest number of variants among the proteins encoded by the *ehxCABD* operon, variants of this protein were investigated further. Of 106 EhxA variants, 99 were likely to be functional and had a total of 166 variable residues. Phylogenetic analysis revealed clustering of variants characteristic of EHEC and aEPEC together (Fig. 6). Moreover, isolates with no pathotype identified and STEC formed two clusters comprising multiple variants. Five variants from nonpathogenic *E. coli* clustered together as a separate group. EhxA subtyping analysis revealed that the subtypes align well with EhxA protein phylogeny and pathotypes (Figs 6 and S9). Additionally, two new subtypes were identified, one of which consisted of five EhxA variants found in nonpathogenic *E. coli*.

Our analyses identified a group of EhxA sequence variants associated with different pathotypes and nonpathogenic *E. coli*, indicating that the functionality of enterohaemolysin as a toxin may be altered in these bacteria.

DISCUSSION

Pathogenic *E. coli* are one of the main bacterial threats to human and animal health [2]. The high degree of genome plasticity leading to a large diversity of virulence factor repertoire among strains is an important factor that limits our ability to design successful preventions and treatment regimens for *E. coli* infections. The development of bacterial genomics in recent decades has allowed better characterization of pathogens, improvement of diagnostics and implementation of new preventive measures to limit disease spread [35]. In this study, we focused on the characterization of the haemolytic *E. coli* population that is associated with multiple *E. coli* pathotypes in human and animal infectious diseases.

Enterohaemolysin was mainly present in intestinal pathotypes such as EHEC, STEC and aEPEC. Similarly to previous studies, we confirmed that enterohaemolysin as a virulence factor could be an important genetic determinant used in routine diagnostics of these pathotypes [36]. STEC infections remain the third most commonly reported zoonosis in the European Union and underreporting remains one of the main issues for public health workers [37]. Therefore, the possible use of enterohaemolysin as an additional marker for diagnostic assays could offer valuable information about the investigated zoonotic agent.

Previous studies focused on enterohaemolysin typing as an additional tool to assess the pathogenic potential of STEC [33, 38]. Six EhxA subtypes based on *ehxA* gene TaqI digestion profile have been described so far, termed A to F [33]. Subtypes B, C and F have been associated with intimin-positive isolates from diseased patients [38, 39]. EhxA subgroups A and D were mainly associated with nonhuman samples and found in intimin-negative isolates. Since our dataset was generated by first searching for enterohaemolysin, we were able to identify a new subtype of EhxA, which was found exclusively in genomes with a VAF profile characteristic of nonpathogenic *E. coli*. The new EhxA subtype clustered with subtype D and was part of a larger clade together with sequences belonging to subtypes D and A (Fig. 6). Our analysis suggests that EhxA sequences from EHEC and aEPEC cluster separately from other EhxA sequences. To date, there have been no studies comparing the biological properties of distinct EhxA variants or subtypes reported. Our analysis provides the framework to address the question as to whether EhxA sequence variation contributes to the biological function of enterohaemolysin, altered *E. coli* toxicity towards the host, and outcome of infection associated with the pathotype definition. Future laboratory analyses should provide an answer to this question and might help with providing proof of the evolution of enterohaemolysin aiding the virulence of human and animal-associated pathogenic *E. coli*.

Alpha-haemolysins or enterohaemolysins are associated with the haemolytic phenotype in *E. coli* strains. We previously reported in a study focused on the characterization of alpha-haemolysin-encoding *E. coli* that the prevalence of these two VAFs does not overlap in the majority of pathotypes [13]. Alpha-haemolysin is mainly associated with UPEC, nonpathogenic *E. coli* or undetermined *E. coli* pathotype, whereas enterohaemolysin is mainly found in pathotypes such as STEC and EHEC. Interestingly, enterohaemolysin or alpha-haemolysin are found in aEPEC isolates. Our data suggest that although both haemolysins belong to the RTX toxin family, they might have a different biological impact on the host. Comparison of biological activity between alpha-haemolysin and enterohaemolysin for various host cells and in different environments could show altered functionality, thereby explaining the distribution of these toxins among *E. coli* pathotypes. Taken together, our analysis reveals a complex relationship between alpha-haemolysin- and enterohaemolysin-positive genomes and diverse *E. coli* pathotypes and provides a framework for understanding the potential role of these VAFs in pathogenesis.

Funding information

R.A.K. was supported by the UKRI Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent project(s) BBS/E/F/000PR10348 and BBS/E/F/000PR10349. M.B. was supported by the Maria Zambrano grant funded by the European Union-NextGenerationEU. K.S. was supported by the National Science Centre grant 2018/31/N/NZ2/01338.

Author contribution

All authors contributed to the manuscript during each step of its preparation (design, experimental procedures, analysis, writing). All authors read and approved the final manuscript.

References

1. Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, *et al*. Recent advances in understanding enteric pathogenic *Escherichia coli*. *Clin Microbiol Rev* 2013;26:822–880.
2. Kaper JB, Nataro JP, Mobley HLT. Pathogenic *Escherichia coli*. *Nat Rev Microbiol* 2004;2:123–140.
3. Köhler C-D, Dobrindt U. What defines extraintestinal pathogenic *Escherichia coli*? *Int J Med Microbiol* 2011;301:642–647.
4. Pizarro-Cerdá J, Cossart P. Bacterial adhesion and entry into host cells. *Cell* 2006;124:715–727.
5. Delicato ER, de Brito BG, Gaziri LCJ, Vidotto MC. Virulence-associated genes in *Escherichia coli* isolates from poultry with colibacillosis. *Vet Microbiol* 2003;94:97–103.
6. Hancock V, Witsø IL, Klemm P. Biofilm formation as a function of adhesin, growth medium, substratum and strain type. *Int J Med Microbiol* 2011;301:570–576.
7. Schiebel J, Böhm A, Nitschke J, Burdukiewicz M, Weinreich J, *et al*. Genotypic and phenotypic characteristics associated with biofilm formation by human clinical *Escherichia coli* isolates of different pathotypes. *Appl Environ Microbiol* 2017;83:e01660-17.
8. Thanassi DG, Nuccio S-P, Shu Kin So S, Bäumler AJ. Fimbriae: classification and biochemistry. *EcoSal Plus* 2007;2.
9. Aleksandrowicz A, Khan MM, Sidorczuk K, Noszka M, Kolenda R. Whatever makes them stick - Adhesins of avian pathogenic *Escherichia coli*. *Vet Microbiol* 2021;257:109095.
10. Garénaux A, Caza M, Dozois CM. The ins and outs of siderophore mediated iron uptake by extra-intestinal pathogenic *Escherichia coli*. *Vet Microbiol* 2011;153:89–98.
11. Richard KL, Kelley BR, Johnson JG. Heme uptake and utilization by Gram-negative bacterial pathogens. *Front Cell Infect Microbiol* 2019;9:81.
12. Caza M, Kronstad JW. Shared and distinct mechanisms of iron acquisition by bacterial and fungal pathogens of humans. *Front Cell Infect Microbiol* 2013;3:80.
13. Kolenda R, Sidorczuk K, Noszka M, Aleksandrowicz A, Khan MM, *et al*. Genome placement of alpha-haemolysin cluster is associated with alpha-haemolysin sequence variation, adhesin and iron acquisition factor profile of *Escherichia coli*. *Microb Genom* 2021;7:000743.
14. Burgos Y, Beutin L. Common origin of plasmid encoded alpha-hemolysin genes in *Escherichia coli*. *BMC Microbiol* 2010;10:193.
15. Bielaszewska M, Aldick T, Bauwens A, Karch H. Hemolysin of enterohemorrhagic *Escherichia coli*: structure, transport, biological activity and putative role in virulence. *Int J Med Microbiol* 2014;304:521–529.
16. Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. Clermontyping: an easy-to-use and accurate in silico method for *Escherichia* genus strain phylotyping. *Microb Genom* 2018;4:e000192.
17. Jolley KA, Maiden MCJ. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 2010;11:595.
18. Ingle DJ, Valcanis M, Kuzevski A, Tauschek M, Inouye M, *et al*. *In silico* serotyping of *E. coli* from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. *Microb Genom* 2016;2:e000064.
19. Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res* 2019;47:5539–5549.
20. Schwengers O, Barth P, Falgenhauer L, Hain T, Chakraborty T, *et al*. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb Genom* 2020;6:mgen000398.
21. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, *et al*. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* 2016;2:e000056.
22. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5:e9490.
23. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47:W256–W259.
24. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, *et al*. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.
25. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;47:D309–D314.
26. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* 2016;17:238.
27. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In: Kotz S and Johnson NL (eds). *Breakthroughs in Statistics: Methodology and Distribution Springer Series in Statistics*. New York, NY: Springer; 1992. pp. 11–28.
28. Galata V, Fehlmann T, Backes C, Keller A. PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res* 2019;47:D195–D202.
29. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, *et al*. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* 2020;21:180.
30. Okonechnikov K, Golosova O, Fursov M, UGENE Team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 2012;28:1166–1167.
31. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;28:3150–3152.

32. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.

33. Lorenz SC, Monday SR, Hoffmann M, Fischer M, Kase JA. Plasmids from shiga toxin-producing *Escherichia coli* strains with rare enterohemolysin gene (ehxa) subtypes reveal pathogenicity potential and display a novel evolutionary path. *Appl Environ Microbiol* 2016;82:6367–6377.

34. Wright ES. Using DECIPHER v2.0 to analyze big biological sequence data in R. *R J* 2016;8:352.

35. Bawn M, Alikhan N-F, Thilliez G, Kirkwood M, Wheeler NE, *et al*. Evolution of *Salmonella enterica* serotype *Typhimurium* driven by anthropogenic selection and niche adaptation. *PLoS Genet* 2020;16:e1008850.

36. Scaletsky ICA, Aranda KRS, Souza TB, Silva NP, Morais MB. Evidence of pathogenic subgroups among atypical enteropathogenic *Escherichia coli* strains. *J Clin Microbiol* 2009;47:3756–3759.

37. Shiga toxin-producing Escherichia coli (STEC) infection. Annual Epidemiological Report for 2019. *Eur Cent Dis Prev Control*; 2021. https://www.ecdc.europa.eu/en/publications-data/shiga-toxin-producing-escherichia-coli-stec-infection-annual-epidemiological [accessed 19 April 2022].

38. Fu S, Bai X, Fan R, Sun H, Xu Y, *et al*. Genetic diversity of the enterohaemolysin gene (ehxA) in non-O157 *Shiga* toxin-producing *Escherichia coli* strains in China. *Sci Rep* 2018;8:4233.

39. Lorenz SC, Son I, Maounounen-Laasri A, Lin A, Fischer M, *et al*. Prevalence of hemolysin genes and comparison of ehxA subtype patterns in *Shiga* toxin-producing *Escherichia coli* (STEC) and non-STEC strains from clinical, food, and animal sources. *Appl Environ Microbiol* 2013;79:6301–6311.